Bioinformatics, 38(2), 2022, 568–569
doi: 10.1093/bioinformatics/btab583
Advance Access Publication Date: 25 August 2021
Applications Note

OXFORD

Systems biology

metID: an R package for automatable compound annotation for LC−MS-based data

Xiaotao Shen [1,†], Si Wu[1,†], Liang Liang[1], Songjie Chen[1], Kévin Contrepois[1], Zheng-Jiang Zhu[2,*] and Michael Snyder[1,*]

[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94304, USA and [2]Interdisciplinary Research Center on Biology and Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, China

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Olga Vitek

Received on May 17, 2021; revised on July 13, 2021; editorial decision on July 31, 2021

Abstract

Summary: Accurate and efficient compound annotation is a long-standing challenge for LC–MS-based data (e.g. untargeted metabolomics and exposomics). Substantial efforts have been devoted to overcoming this obstacle, whereas current tools are limited by the sources of spectral information used (in-house and public databases) and are not automated and streamlined. Therefore, we developed metID, an R package that combines information from all major databases for comprehensive and streamlined compound annotation. metID is a flexible, simple and powerful tool that can be installed on all platforms, allowing the compound annotation process to be fully automatic and reproducible. A detailed tutorial and a case study are provided in Supplementary Materials.

Availability and implementation: https://jaspershen.github.io/metID.

Contact: mpsnyder@stanford.edu or jiangzhu@sioc.ac.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Liquid chromatography coupled to mass spectrometry (LC−MS) is a comprehensive, unbiased technology to research small compounds, which has become increasingly popular in food, environment and biomedical studies (Fraga-Corral *et al.*, 2020; Wishart 2016). This approach has been very successful however the generated data has typically not been used to its full potential in part due to an incomplete annotation of the chemicals detected. Implementing an accurate and efficient compound annotation workflow would be invaluable to assist biological hypothesis generation and data interpretation. In order to standardize data reporting from metabolomics studies, the community proposed to adopt a grading scheme for annotation confidence ranging from 1 to 4 (Sumner *et al.*, 2007) using the following parameters: mass to charge ratio (*m*/*z*), retention time (RT) and MS^2 spectral matching. Confident annotation being the Achilles' heel of modern metabolomics experiments, recent initiatives are now focusing on generating high-quality MS^2 spectral libraries that are publicly accessible. By taking advantage of these resources, some tools (Chaleckis *et al.*, 2019) were developed to facilitate compound annotation. However, these tools are limited by: (i) the breadth of parameters used to compute an annotation score; previous tools typically only use *m*/*z* and/or MS^2 spectra information but lack RT; (ii) the inability to combine spectral information from multiple sources including in-house and public databases in a streamlined fashion; (iii) the processing speed, previous tools normally do not allow the implementation on cluster servers. More detailed comparisons between the existed tools and metID are listed in Supplementary Table S1.

In this context, we developed a new R package, metID, particularly designed to (i) streamline the construction of users' in-house databases from authentic chemical standards run in each laboratory and (ii) automated compound annotation pipeline. As annotation level 1, metID provides our spectral and RT data from our in-house database containing more than 1,000 authentic standards acquired in HILIC and RPLC modes (users need to use the same LC gradient to match RT). The users can also use the metID to construct their own in-house database for level 1 annotation. For level 2, metID now provides five public MS^2 databases, and five MS^1 databases for level 3 (Supplementary Table S2).

2 Features and methods

Using metID, users can easily build their in-house databases using the data from authentic standards that were acquired in their own laboratories. The public databases can also be easily organized as the database format for metID. The in-house databases in our laboratory and several public databases are provided as reference. metID can be installed on all platforms (Mac OS, Windows and Linux).

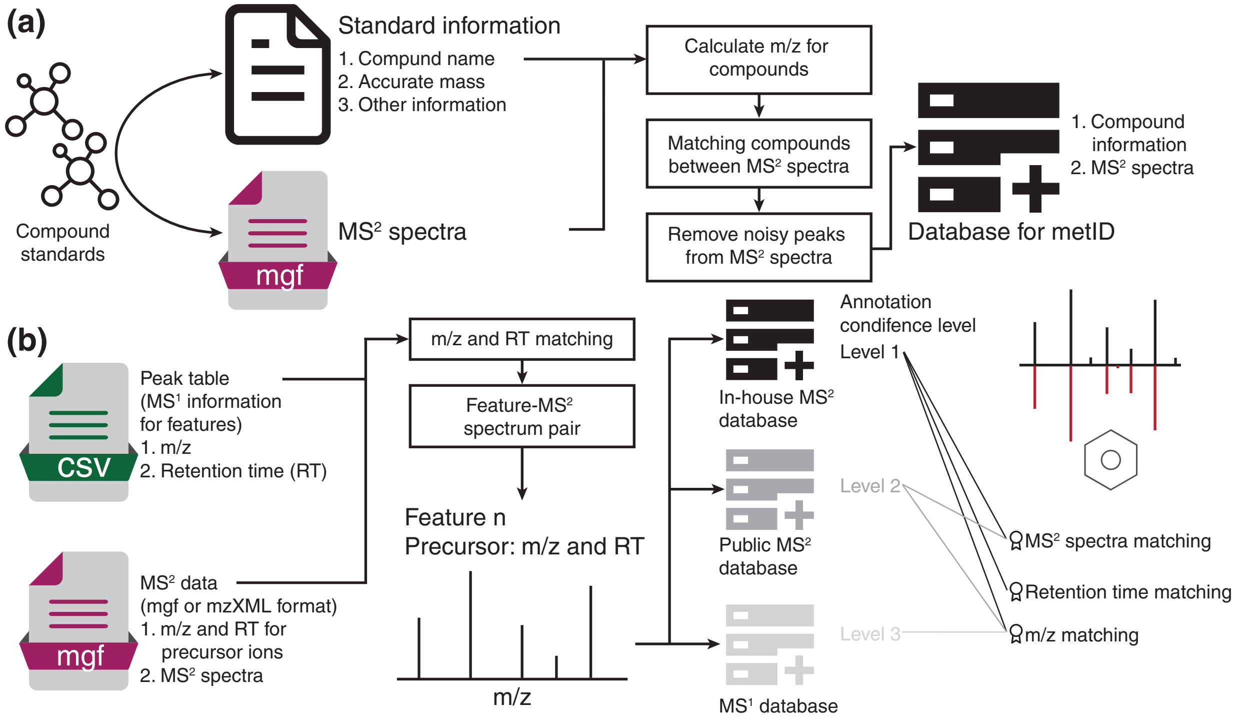

Fig. 1. The overview of (**a**) in-house database construction and (**b**) compound annotation in different levels

2.1 Database construction for metID

The users can use the in-house library embedded in metID, which requires the users to use the exact same LC/MS instrumental setting as metID provided. If users have in-house standards which have been acquired with RT and MS^2 spectra, it is possible to build their own in-house database using the 'construct_database()' function (Fig. 1a). Three items are contained in the database: (i) database information, (ii) standard information and (iii) MS^2 spectrum for each compound. If the users want to use the in-house library embedded in metID but without RT information, they can set the 'rt.match.tol' beyond your LC gradient time range.

2.2 Compound annotation

RT may shift in different acquisition batches. Therefore, it is necessary to correct the RT in databases using the 'correct_database_rt()' function if you spike internal standards into standards and subject samples. Then metID can annotate compounds with different levels according to databases (in-house database, level 1; public MS^2 database, level 2; MS^1 database, level 3). For *m/z*, RT and MS^2 spectra match scores, they are combined as one total score and scaled to 0–1 (Shen *et al.*, 2019; Tsugawa *et al.*, 2015) (Fig. 1b and Supplementary Material).

3 Case study

We applied metID on a published study from our lab (Contrepois *et al.*, 2020) as a case study to demonstrate the value of metID for automatic metabolite annotation (only ~2 h running on a Mac with 6-core 32 G memory). The authors reported 463 annotated metabolites with level 1–2 by manual inspection of *m/z*, RT and MS^2 (Level 1 as golden standards). By using metID, we successfully annotated 942 metabolites with high confidence (level 1–2). Comparing the annotation results from metID with the original annotation in the publication, metID retrieved all the 463 annotated metabolites in the previous paper in an automatic way. Besides these overlapping annotations, metID can annotate 479 new metabolites that were not annotated in the original publication (Supplementary Fig. S5). These results indicate that metID is not only a valid approach for compound annotation with high accuracy (100%) and high speed (~2 h) but also a powerful tool to largely increase annotation coverage in an automatic way (Supplementary Material).

4 Conclusion

metID is used for compound annotation in untargeted LC–MS-based data (metabolomics and exposomics). It allows users to build and share their in-house databases and search against all the major spectral databases. To our best knowledge, it is the first R package to allow users to build in-house databases and combine them with the public databases for compound annotation and can be applied in all computer platforms automatically. As R is very popular in the bioinformatics field, it means that as an open-source tool, metID can contribute to all community members to increase its public databases and add new methods and steps to it.

Financial Support: none declared.

Conflict of Interest: M.S. is a co-founder and member of the scientific advisory boards of the following: Personalis, SensOmics, Filtricine, Qbio, January, Mirvie and Oralome.

References

Chaleckis,R. *et al.* (2019) Challenges, progress and promises of metabolite annotation for LC–MS-based metabolomics. *Curr. Opin. Biotechnol.*, **55**, 44–50.

Contrepois,K. *et al.* (2020) Molecular choreography of acute exercise. *Cell*, **181**, 1112–30.e16.

Fraga-Corral,M. *et al.* (2020) Analytical metabolomics and applications in health, environmental and food science. *Crit. Rev. Anal. Chem.*, **10**, 109.

Shen,X. *et al.* (2019) Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. *Nat. Commun.*, **10**, 1516.

Sumner,L.W. *et al.* (2007) Proposed minimum reporting standards for chemical analysis. *Metabolomics*, **3**, 211–221.

Tsugawa,H. *et al.* (2015) MS-DIAL: data-independent ms/ms deconvolution for comprehensive metabolome analysis. *Nat. Methods*, **12**, 523–526.

Wishart,D.S. (2016) Emerging applications of metabolomics in drug discovery and precision medicine. *Nat. Rev. Drug Discov.*, **15**, 473–484.